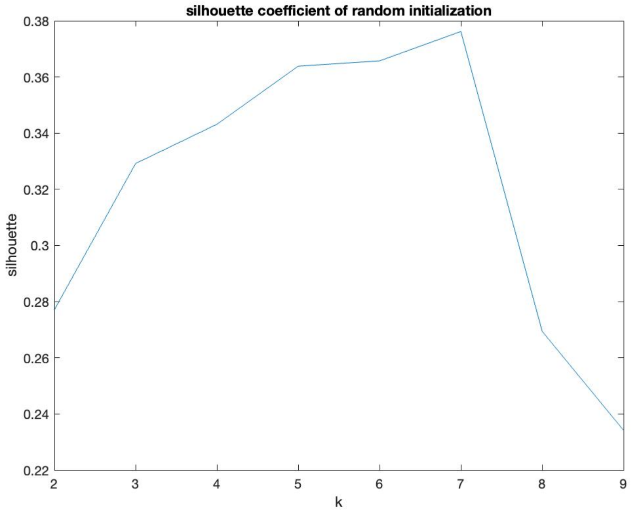
Task 2:

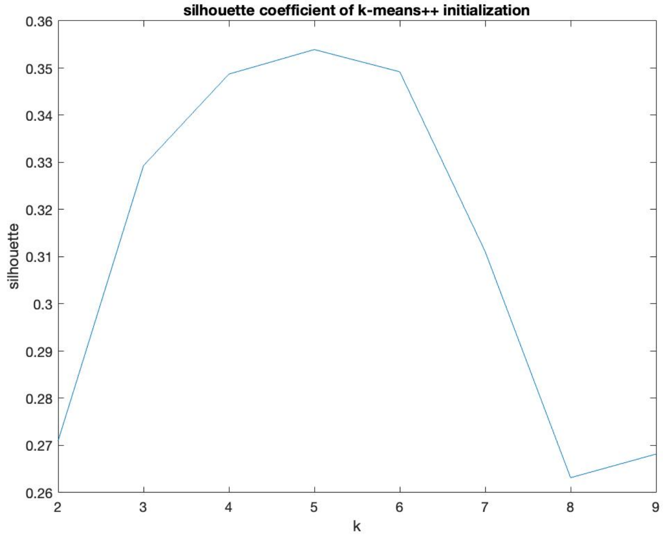
k	silhouette coefficient
2	0.27700962142660707
3	0.3291854105629588
4	0.3431492854250219
5	0.36384782794374604
6	0.36573063831994057
7	0.3761998366998852
8	0.269422407514708
9	0.23419786477260615



Using random initialization, the silhouette coefficient for different clusters ranging from 2 to 9 is shown in the table and figure above. With the highest silhouette coefficient, the best k is 7.

Task 3:

k	silhouette coefficient
2	0.270942553577313
3	0.3293021174466361
4	0.3487195884380835
5	0.3539202709858661
6	0.34920944719518315
7	0.31110567988668263
8	0.263169633318491
9	0.26817416812127365



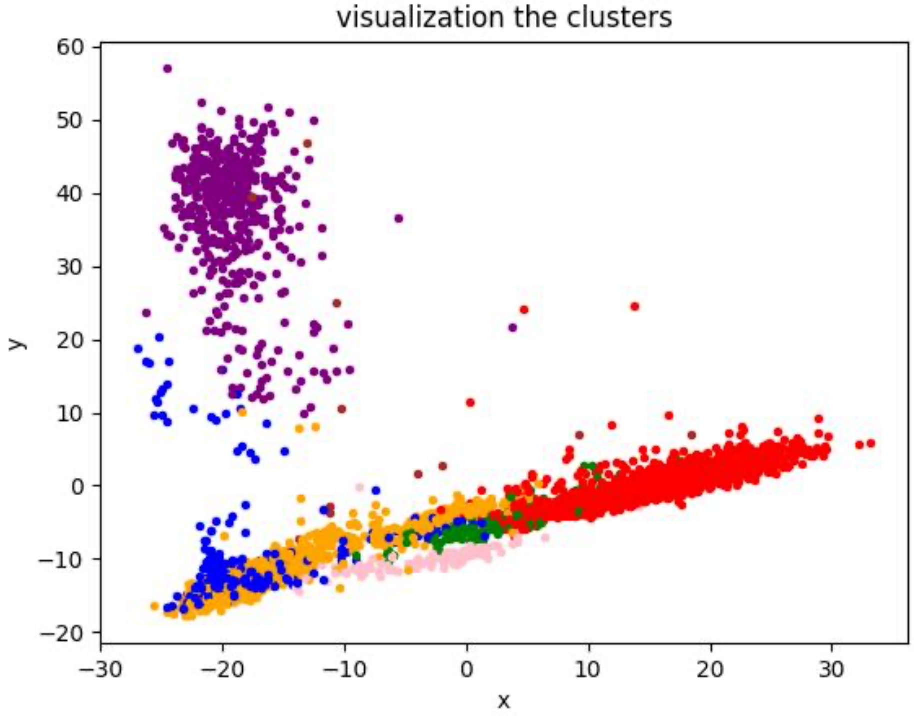
Using k-means initialization, the silhouette coefficient for different clusters ranging from 2 to 9 is shown in the table and figure above. With the highest silhouette coefficient, the best k is 5.

The silhouette coefficient for the random initialization increases with k before k is 7 and decreases with k after k is 7. Because of the random initialization, the outcome clusters with the same k may differ. Also, this algorithm is sensitive to outliers. If the k is small, there may be an undesirable cluster containing the outlier. But, if the k is large, the outlier may become one cluster. However, large k is less friendly to more aggregated data, and data that are close enough to each other are divided into different clusters.

The silhouette coefficient for the kmeans++ initialization increases with k before k is 5 and decreases with k after k is 5. This algorithm tends to choose the outliers (the points that has the longest distance) be the centroids when initialize. When an outlier is chosen and the k is small, the silhouette coefficient will be relatively large. For example, k is 2 and the silhouette coefficient is 0.55 (when the outlier is chosen). it shows that the kmeans++ is fit for wide range dataset, and it is less sensitive to the outliers.

All in all, these two algorithms apply to different situations. The random initialization is suitable for more aggregated datasets, and the kmeans++ is suitable for more dispersed datasets which has outliers.

Task4:


visualization the clusters
y
x